Error detection and error correction for improving quality in machine translation and human post-editing

Lucia Comparin[a,b], Sara Mendes[a,c]

[a] Universidade de Lisboa, Centro de Linguística da Universidade de Lisboa
Alameda da Universidade, 1600-214 Lisboa, Portugal
[b] Unbabel, Rua Visconde de Santarém, 67B, 1000-286 Lisboa, Portugal
[c] Faculdade de Letras da Universidade de Lisboa
Alameda da Universidade, 1600-214 Lisboa, Portugal
lcompa@gmail.com, s.mendes@campus.ul.pt

Abstract. Machine translation (MT) has been an important field of research in the last decades and is currently playing a key role in the translation market. The variable quality of results makes it necessary to combine MT with post-editing, to obtain high-quality translation. Post-editing is, however, a costly and time-consuming task. Additionally, it is possible to improve the results by integrating more information in automatic systems. In order to improve automatic systems performance, it is crucial to evaluate the quality of results produced by MT systems to identify the main errors. In this study, we assessed the results of MT using an error-annotated corpus of texts translated from English into Italian. The data collected allowed us to identify frequent and critical errors. Detecting and correcting such errors would have a major impact on the quality of translation and make the post-editing process more accurate and efficient. The errors were analyzed in order to identify patterns of errors, and solutions to address them automatically or semi-automatically are presented. To achieve this a set of rules are formulated and integrated in a tool which detects or corrects the most frequent and critical errors in the texts.

Keywords: machine translation, human post-editing, error detection, rule-based editing

1 Introduction

Machine translation (MT) has been an important field of research since the second half of the 20th century. The work done in the area enabled improvements in the results, and the development of different systems, while simultaneously encouraging the work in related areas, such as computational linguistics and machine learning. Thanks to research in these fields and to the improvements achieved, MT has become an important part of the translation process in the current market, as it plays a key role in handling the increasing volume of translation needed and the short time available to deliver it. Despite the increasing use of MT in the translation market, the quality of the results is still variable and dependent on several aspects such as the paradigm of

the MT system used and the intended use of MT (Dorr et al, 1999: 36). Additionally, the MT systems currently available are numerous and their performance is not alike in terms of quality. These two aspects, namely the variability of the results and the number of different types of MT systems, make the evaluation of the systems a necessary step not only to accurately characterize different MT systems and their performance, based on the quality of results, but also to define how MT systems can be improved.
In this paper, we take data from a quality assessment task of MT results to outline strategies to tackle the most frequent errors identified. Quality assessment allows not only to understand whether a MT system produces satisfactory results, but also to identify the aspects that have to or can be improved. The work presented here focuses on the second aspect and it has been carried out in collaboration with Unbabel, a startup company that offers almost real-time translation services, combining MT with crowd post-edition.

2 Related Work

The wide adoption of MT systems both for gisting purposes and to produce professional quality translations, has generated the need for methods to evaluate the performance of MT systems and assess the results.
There is extensive work in describing MT errors, usually involving post-hoc error analysis of specific MT systems (e.g. Kirchhoff et al., 2007, Vilar et al., 2006) rather than online error detection. One exception is Hermjakob et al. (2008), who studied named entity (NE) translation errors, and integrated an improved on-the-fly NE transliterator into a statistical machine translation (SMT) system. Several error taxonomies have been created in order to classify errors (e.g. Vilar et al, 2006, Popović and Buchardt, 2011). A linguistically motivated error taxonomy has been presented by Costa et al. (2015) to classify translation errors from English into European Portuguese, while Lommel (2015) outlined a framework to develop a customized error taxonomy, under the scope of the Quality Translation 21 project.
In the evaluation process of MT systems, error taxonomies are used both in human and automatic annotation. The former, performed by a human annotator, involves a process of annotation and analysis as described in Daems, Macken & Vandepitte (2014) and Stymne & Ahrenberg (2012), while the latter presupposes a metric to automatically assess quality in machine translated texts, using a human translation as a reference. The most widely used are BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014).
Apart from evaluating the performance of the system, in some cases, such as when MT is used in the translation market, it is necessary to predict the quality of a translation. This practice is referred to as Quality Estimation (QE), which differs from standard MT evaluation for not having access to human translations as a reference for the evaluation and not being performed by a human evaluator. Among the numerous applications of QE we underline the ability to decide which segments need revision by a translator (quality assurance and error detection).
Work in QE for MT started in the early 2000's, as an analogy to the confidence scores used in Speech Recognition, which essentially consisted on the estimation of word

posterior probabilities. Presently, QE aims at estimating more interpretable metrics, which have been used in many different tasks, such as improving post-editing efficiency by filtering out low quality segments (Specia et al., 2009; Specia, 2011), selecting high quality segments which do not require post-editing (Soricut and Echihabi, 2010), or highlighting sub-segments that need revision (Bach et al., 2011).
In this study we aim at developing an approach that is able to aid human post-edition of machine translated texts from English to Italian. To achieve this, we used a corpus of machine translated texts from English into Italian to identify repeated patterns in the errors. This allowed us to identify generalizations and outline strategies to address the errors automatically or semi-automatically: a set of rules was formulated and integrated in the Smartcheck, a tool developed by Unbabel that checks format, grammar and style in the MT texts. The Smartcheck analyzes the translated segments and underlines the expressions where an error is identified, providing suggestions to address it to the human editor.

3 Methodology

For the study discussed in this paper we considered the language pair English-Italian. As a starting point of our work, we used a manual error annotation of a corpus of translated texts. The texts in the corpus were translated using the Google Translator Translation API, as this was the MT system initially being used at Unbabel. Google Translator is a free SMT system available online that uses a data-driven approach based on web content, currently allowing to translate from and into more than 70 languages. More recently, a MT system for the English-Italian language pair was trained using Moses (Koehn et al., 2007). Moses is an open-source SMT system that enables the user to train translation models for any language pair. It is composed of a training pipeline, which comprises the stages involved in the translation process, such as tokenization, alignment, acquiring a language model, and automatically selecting the best possible translations among the results of different statistical models, and a decoder. The advantage of using Moses is that it can be customized to the needs of the user, that the text does not go through an external server, and that additional tools can be integrated. For this reason, and for evaluating how dependent on the MT system used to produce the data analysed our proposal is, we tested the impact of the rules proposed on error reduction and error detection on machine translated data produced by Google Translator and Moses.
In order to formulate the generalizations allowing for an automatic detection and/or correction of translation errors, we identified and categorized the most common errors occurring in an annotated corpus of MT outputs. Once these errors were collected, we conducted an analysis of the data which allowed us to find error patterns, from which we outline a set of rules to automatically address the specific issues detected as shortcomings of Google Translator.
There are several difficulties in trying to establish rules that can improve translation results, the main being the ability to formulate rules that can be applied to each case observed to solve the issue, without introducing problems in other examples, i.e. without over generating. Another important aspect is related to whether the generali-

zation identified involves only the target text or needs to be aware of the relation between the target structure and the source text that was translated, which makes implementation more challenging.
In this paper, we propose two types of rules: rules for error correction and for error detection. The motivation for doing so relies on the fact that many generalizations involve complex linguistic phenomena such as ambiguity which require human inspection to guarantee the quality of the results. Additionally, the context in which the work presented here has been developed requires high quality of the results and aims at aiding human post-edition of MT results, which is entirely in line with a semi-automatic approach such as the one associated to the second set of rules.
To present our approach and the results obtained, we discuss in detail two types of error, identifying generalizations in terms of what is problematic for Google Translator, putting forth a small set of rules to tackle the shortcomings identified and showing their impact in terms of automatic error reduction and error detection.

4 The error annotated corpus

Error annotation consists of the identification, categorization and analysis of errors in a text. In this study, we used a manually annotated corpus. Human annotation, on one side, is expensive and time consuming. It is also more difficult to achieve consistency and objectiveness when the annotator is human. On the other side, it is more accurate and can provide a more thorough analysis of the errors.
The corpus was annotated using an in-house tool to assess the quality of the texts delivered to clients in different language pairs. This annotation tool shows the source text, the target text, warnings automatically produced by a checker, and glossary terms. Errors were annotated in the target text and then classified according to an error typology developed at Unbabel, following the documents and general guidelines of the Multidimensional Quality Metrics (MQM) framework (Lommel, 2015) and TAUS documents (www.taus.net).
This error typology consists of 41 error categories that are included in 7 major categories (see Table 1). Only leaf entries of the taxonomy tree could be selected and marked as errors in the annotation task. In the specific case of the annotation considered in this paper, only one category could be selected for each expression, which amounts to say that while the most critical or relevant errors are marked others that do occur in the corpus were not annotated. The criteria for marking an error instead of another were based on the impact each of them has on the quality of the machine translated text. For details regarding the error typology considered and the specifications of the annotation task see Comparin (2016).
For the work presented in this paper we randomly took a corpus of 50 texts translated from English into Italian with a SMT system, Google Translator, reviewed by human editors and annotated both after MT and after the first post-edition. The size of the annotated texts included in the corpus ranged from 100 to 700 words and covered domains such as tourism, client support and e-commerce. The reason for analyzing the machine translated text was to study the errors present, categorize them, and try to solve the most critical and most recurrent ones. The annotation of the first human post-edition is useful to provide us with information about the errors the editors cor-

rect and those that persist along the different stages of the translation process. In the context of the work presented here we only considered the errors after MT, as our goal consists on evaluating how many errors can be avoided by improving the automatic tools operating on the text after MT, either by automatically correcting the errors or by providing more precise and useful warnings to human editors.

TYPE OF ERROR	MT	FIRST POST-EDITION
Accuracy errors	236	55
Fluency errors	848	83
Style errors	1	3
Terminology errors	0	14
Wrong language variety errors	0	0
Named entities errors	19	15
Formatting and encoding errors	0	0
Total	1.104	170

Table 1. Total number of errors annotated in the corpus per general error category

FLUENCY ERROR TYPES	MT	FIRST POST-EDITION
Word selection	1	1
Tense selection	0	0
Coherence	2	1
Duplication	0	0
Orthography	1	1
Capitalization	52	19
Diacrits	0	0
Punctuation	9	4
Unpaired quote marks and brackets	1	0
Whitespace	17	5
Inconsistency in character use	0	0
Function words	0	0
Prepositions	70	10
Conjunctions	12	1
Determiners	237	19
Part-of-speech	30	1
Agreement	159	13
Tense/mood/aspect	101	3
Word order	106	4
Sentence structure	50	1
Total	848	83

Table 2. Distribution of fluency errors annotated in the corpus per specific categories considered in the taxonomy

In Table 1, we present the error annotation data used in this study. As we can see, the number of errors annotated is high and is not evenly distributed among the different

categories. This is certainly not independent of the fact that only the most relevant or critical error was marked when there was more than one error in a word or phrase.
In Table 2, above we present fluency errors, the general category with the greatest amount of errors, in more detail. Among these, the category with the highest number of errors annotated, in machine translated texts, is "determiners", followed by "agreement", "tense/mood/aspect", and "word order". These include errors that can prevent the reader from understanding the text clearly, having a major or critical impact on the quality of the translation. Due to space limitations, in this paper we will focus on the discussion of two of these error types: "word order" and "agreement".

5 Word order: noun modification structures

Word order is a crucial aspect of language, often playing a decisive role in the grammar of specific languages. In this study, we decided to focus on word order errors in noun modification structures, as these were by far the most representative in our corpus (see Table 3). Moreover, the issue is crucial because, even if there are errors of this type in which the editor easily and quickly understands the correct word order, some of the translated structures are ambiguous, leading the editor to spend a considerable amount of time to produce the correct translation structure in the target language.

WORD ORDER ERRORS	
Number of unique errors	68
Number of unique errors in noun modification	65
Number of unique errors involving other structures	3
Total number of annotated errors	106

Table 3. Distribution of word order errors occurring in the corpus per type of structure

There are different ways of modifying a noun in languages like English: with an adjective; a past participle; a prepositional phrase; another noun; a verb in the –ing form; a relative clause. Moreover, a noun can have more than one modifier (e.g. 'the old leather sofa', 'my blue scarf with dots', 'the dish left on the table').

5.1 Word order errors in the corpus

The 65 unique errors involving a noun modification structure were divided into the subcategories considered in Table 4. In addition to these distinctions, we noticed that, in 27 of the 65 errors occurring in noun modification structures, the head noun of the NP is a NE: 17 cases occurring in NN modification structures, 1 in an adjective-noun modification structure, and 9 in modification structures with both adjectives and nouns. The high occurrence of errors annotated involving NE highlights how challenging these constituents are to MT. Besides their idiosyncratic behaviour in language, NE are often not included in lexical resources due to their low frequency of occurrence, and are often not present in the corpora used to train MT systems for the same reason.

WORD ORDER ERRORS IN NOUN MODIFICATION	
Errors in noun-noun modification (NN modification)	29
Errors in adjective-noun modification	4
Errors in noun modification with both noun(s) and adjective(s)	32
Total	65

Table 4. Distribution of word order errors in noun modification structures considering different types of constructions

Word order errors involving noun modification structures were seen in NN modification structures in English which can be translated by two different structures in Italian: a PP or an ADJP. Noun modification with both noun(s) and adjective(s) errors include both the cases in which all the constituents modify a single head noun, and the cases of inlaid modification, in which one or more modifying constituents modify a modifier of the head noun. With regard to adjective-noun modification errors, for which only 4 occurrences are found in our corpus, we can state that these are in general correctly dealt with by the system.

5.2 Tackling word order errors in noun modification structures

Considering the specific case of modification structures, there are often problems when the head noun has more than one modifier. In such cases, the order in the translation is critical because it can be an inlaid modification structure and, therefore, a modifying noun or an adjective modify another modifier, and not the head noun. In such sequences of constituents, dependency is often different from case to case.
Based on our corpus study, we established the following set of rules to address word order errors. As many of the structures can be ambiguous, we decided to implement most of the rules in the checker, for posterior human inspection.

Rules for error detection by the checker.

Rule 1

When a named entity occurs in the target text and is preceded or followed by an adjective or a PP that modifies it, ask the editor to check the order of the elements in the sentence.

(ADJP|PP)+PROPN → warning
PROPN+ (ADJP|PP) →warning

Rule 2

When a named entity occurs in the target text within a PP as a modifier, ask the editor to check the order of the elements in the sentence.

$N+^{modifies}P + PROPN \rightarrow$warning

Rule 3

If a noun or a PP precede the head noun, ask the editor to check the order of the elements in the sentence.

$(N|PP)+N \rightarrow$ warning

Rule 4

If one of the sequences listed below are detected, ask the editor to check the order of the elements in the sentence.

$N+N^{+} \rightarrow$ warning
$N+ADJ^{+}+N \rightarrow$ warning
$ADJ^{+}+N+N \rightarrow$ warning
$ADJ +ADJ^{+}+ N+N^{+} \rightarrow$warning

Rules for error correction.

Rule 5

If there is an adjective modifying a noun in English and the adjective is a quality adjective, then the order in the target language should be noun adjective[1].

$ADJ_{Q}+N \rightarrow N + ADJ_{Q}$

Rule 6

If there is a noun preceding another noun in English, and the first noun modifies the second, invert the order and convert the noun into an adjective phrase or a PP.

$N1+^{modifies}N2 \rightarrow N2+(ADJP|PP_{N1})$

This second set of rules requires a tool that checks both the source and the target text and introduces changes, i.e. corrects, MT results automatically. In this case it is particularly important that such rules do not overgenerate, which is why most of the rules proposed are implemented in the checker to aid post-edition. In section 7 we evaluate the coverage and impact of the rules proposed both for addressing word order errors and agreement errors, which are discussed in the next section.

6 Agreement

Agreement, in linguistics, is the morphosyntactic covariation of two or more words in a sentence. Agreement is a complex issue in MT and a frequent source of error. Errors can occur both in the analysis of the source text or in the generation of the target text. In the former, the system, in case of error, does not extract relevant information about

[1] The correct implementation of this rule would require a rich lexical resource providing the classification of adjectives, which is not currently possible with the resources available at Unbabel and for the language pair considered.

agreement features in the source text and is therefore unable to provide crucial information to the module which generates the translation. In the latter, although the system extracts the correct information regarding the relevant agreement features in the source text, it is unable to generate the correct output in the target language.
There are various aspects which make agreement a challenging phenomenon to be handled by automatic systems such as MT. The first difficulty is the fact that a word can have contrasting agreement features in the source and target languages. The system must have this information and for that it needs access to a rich lexical resource. These are expensive and require time and effort to be always updated and complete. Additionally, the lexicon is open and constantly changing, making it difficult to achieve completeness and accuracy in lexical resources.
The second difficulty amounts to the fact that the source and target languages can have contrasting morphological systems, one being richer than the other, as in the case of Italian and English, since Italian has a richer inflectional morphology than English. This is particularly problematic, when the target language has a richer inflectional morphology than the source language, as the system does not find in the source text the information needed for the generation of the target text.
Another difficulty consists of assessing the correct dependency between constituents in long or complex sentences. The structure of a noun phrase can be ambiguous due to the position of the constituents. This happens when a word can agree with more than one word co-occurring with it. This situation is even more problematic when the constituents are not morphologically marked in the source language, for example. Another difficulty, which has already been mentioned in the previous section, involves NE, which have a very idiosyncratic behaviour, which is often contrasting in the source and target languages, besides not being encoded in lexical resources.

6.1 Agreement errors in the corpus

There are many agreement errors annotated in the corpus. On one side, the errors from this category are apparent, and the editors usually correct them without spending too much time in this task. On the other side, agreement errors are common and, if the editor happens not to notice one of them, they cause the text to be looked at as a sloppy translation. This is the reason why agreement errors are considered severe even if they do not hinder access to the content of the text. Given all this, it is useful to automatize the agreement error correction and detection as much as possible.

AGREEMENT ERRORS		
	annotated in the corpus	detected by the checker
Gender agreement errors	137	51
Number agreement errors	19	0
Person agreement errors	3	0
Total	159	51

Table 5. Agreement errors in the corpus and detected by the checker

Among the 159 agreement errors identified in the annotation task (see Table 5), only approximately 1/3 were already detected by the checker. As we can see in Table 5, all

the agreement errors detected by the checker were gender agreement errors and they were all agreement errors between the determiner and the head noun of an NP.

6.2 Tackling agreement errors

Agreement errors can be avoided if the syntactic dependency among the constituents in a phrase is correctly identified. In order to do so, we used a parser (Martins et al., 2013) to analyze dependencies in the target text, which was able to correctly identify the value for agreement features in the majority of the cases in our corpus, even when there were agreement errors. In such cases the parser was able to correctly identify the value for the relevant agreement features for the separate constituents in most of the cases. Assessing the performance of the parser in the analysis of incorrect sentences or phrases is crucial in understanding which solutions are possible to implement and, on the contrary, which have to be addressed using different strategies.
Morphological regularities in the grammar of Italian allows us to come up with some rules for the checker.

Rules for error detection by the checker.

Foreign words in Italian are in general masculine. Even if the parser does not identify whether a given noun is a foreign word or not, no Italian noun ends in a consonant. Therefore it is possible to say that if a noun ends in a consonant, it is a foreign word.[2] Also, since in Italian there are no words ending in "–s", we can have the checker highlight all the words with this ending found in the target text, first controlling whether they are NE, that are in general correctly tagged as PROPN by the parser, this way identifying foreign nouns in the plural form, which are incorrect in Italian.

Rule 7

If a noun ending in a consonant occurs in the target text, check if its specifiers and modifiers are masculine.

$$\mathrm{SPR}^{*} + \mathrm{N}_{-\mathrm{consonant}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\mathrm{masc}} + \mathrm{N}_{-\mathrm{consonant}} + \mathrm{MOD}^{*}{}_{\mathrm{masc}}$$

Rule 8

If a noun ending in an -s occurs in the target text, check if it is a foreign word in plural form.

As we already mentioned, the parser classifies NE as proper nouns, but does not identify their agreement features. Therefore, the only solution to address errors involving this type of nouns would be highlighting them when they have dependency relations with other elements in the sentence. Naturally, another possibility would involve en-

[2] Please note that the rule presented does not cover all the cases in which a foreign word occurs, since there are foreign words ending in a vowel, such as "cookie", although it significantly reduces the number of errors.

rich lexical resources with information regarding these lexical items, which is out of the scope of this paper.

Rule 9

When a named entity occurs in the target text co-occurring with specifiers and modifiers, ask the editor to check the agreement between all these elements.

SPR*+MOD*+PROPN+MOD* → warning

Rule 10

If the quantifier "nessuno" or "chiunque" are part of the subject of a sentence, ask the editor to check if the head verb form of the sentence is singular.

Rules for error correction.

Rule 11

If a noun ending in "–tore" occurs in the target text, then its specifiers and modifiers are masculine

$$\mathrm{SPR}^{*} + \mathrm{N}_{-\mathrm{tore}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\mathrm{masc}} + \mathrm{N}_{-\mathrm{tore}} + \mathrm{MOD}^{*}{}_{\mathrm{masc}}$$

Rule 12

If a noun ending in "–tà", "–tù", "–trice", "–tite" or "-zione" occurs in the target text, then its specifiers and modifiers are feminine.

$$\mathrm{SPR}^{*} + \mathrm{N}_{-\mathrm{tà}|-\mathrm{tù}|-\mathrm{trice}|-\mathrm{tite}|-\mathrm{zione}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\mathrm{fem}} + \mathrm{N}_{-\mathrm{tà}} + \mathrm{MOD}^{*}{}_{\mathrm{fem}}$$

7 Results and final remarks

To assess the impact of the rules proposed above in error reduction and error detection, we tested them against two comparable annotated corpora analogous to the corpus used in the data study described in previous sections, each with 50 machine translated texts, one with outputs from Google Translator and the other with translation results obtained with the SMT trained with Moses. As mentioned earlier, the SMT trained using Moses has only recently integrated in Unbabel's workflow. Since then Unbabel is focusing on quality in Help Center emails and is therefore annotating only this type of texts. Considering this, and for the sake of obtaining comparable results we also used only this type of text for the Google Translator testing data. This was not the case, however, of the corpus used for the data study, which also included texts related to tourism, that are usually longer and involving more modification structures than support emails. This results in a clearly contrasting amount of annotated errors - the number of errors in the first corpus was much higher than that observed in the two corpora of the second annotation -, even if all three corpora had roughly the same size. In Tables 6 and 7, we can see that the texts translated using Google SMT contained less errors than those translated using the company's Moses SMT. The differ-

ence is particularly relevant in the case of agreement errors. It is also apparent that the rules proposed to address word order errors have a significantly higher coverage than those formulated to tackle agreement errors.

	Google Translator		Moses SMT	
	#	%	#	%
errors corrected	3	8.6%	1	2.5%
errors detected by the checker	27	77.1%	23	57.5%
errors not covered by the rules	5	14.3%	16	40%
Total number of errors	35	100%	40	100%

Table 6. Word order errors covered by the rules proposed in texts translated by two different SMT

	Google Translator		Moses SMT	
	#	%	#	%
errors corrected	4	7.8%	-	0%
errors detected by the checker	9	17.7%	7	10%
errors previously covered	5	9.8%	8	11.4%
errors not covered by the rules	33	64.7%	55	78.6%
Total number of errors	51	100%	70	100%

Table 7. Agreement errors covered by the rules proposed in texts translated by two different SMT

In fact, a significant amount of agreement errors (9 in Google Translator outputs and 24 in Moses's) involved a constituent of the VP and no rule was suggested to solve such errors. Since for the sake of this work we have tested the performance of the parser on ill-formed structures regarding local agreement with suitable results for our goals, as future work we will seek to extend our approach to outline possible strategies for tackling agreement errors involving constituents in the VP, which typically involve longer distance and more diverse syntactic relations.

Still regarding the specific case of agreement errors and rules to identify them, there are many agreement errors occurring in a sequence of a determiner and a head noun ending in –o that still are not covered by the rules. Nouns with this ending in Italian are masculine in the majority of the cases. There are many exceptions, however, which is why these have not been covered by the rules.

As made apparent by tables 6 and 7, there is a clear effect on the results when the MT system used changes. On the one hand, because Moses SMT has a poorer performance for the two types of error considered, but also because it generates errors that were not originally seen in the first corpus analyzed. Also, the performance with agreement errors has a poor coverage, even if we consider that the maximum coverage would be around 75% of the data, as no rules were formulated for addressing errors observed within the VP. In this case, a new iteration of our approach with the test corpora will be developed to understand the kind of structures that are not being grasped by our rules and outlining additional rules to address them.

Finally, and considering our approach has a greater impact on error detection than automatic error correction, as future work we will also test the impact of the checker warnings on error reduction and efficiency improvement of human post-edition.

8 References

1. Costa, Â., Ling, W., Luís, T., Correia, R., Coheur, L. (2015). "A linguistically motivated taxonomy for Machine Translation error analysis", *Machine Translation*, *29*(2), pages 127-161.
2. Bach, N., Huang, F., Al-Onaizan, Y. (2011) "Goodness: a method for measuring machine translation confidence", *Proceedings of ACL11*, pages 211–219, Portland, USA.
3. Comparin, L. (2016) *Quality in Machine Translation and Human Post-editing: Error Annotation and Specifications*, MA thesis, Universidade de Lisboa.
4. Daems, J., Macken, L., Vandepitte, S. (2014, May). "On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship", *LREC*, pages 62-66.
5. Dorr, B. J., Jordan, P. W., & Benoit, J. W. (1999). "A survey of current paradigms in machine translation", *Advances in computers*, *49*, pages 1-68.
6. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007) "Moses: Open Source Toolkit for Statistical Machine Translation", *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.
7. Lavie, A., Denkowski, M. J. (2009). "The METEOR metric for automatic evaluation of machine translation", *Machine translation*, *23*(2-3), pages 105-115.
8. Lommel, A. (2015) *Multidimensional Quality Metrics MQM Definition*.
9. Martins, A., Almeida, M., Smith, N. (2013) "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers." in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, August 2013. (Available at: https://www.cs.cmu.edu/~afm/Home_files/acl2013short.pdf).
10. Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002) "Bleu: a method for automatic evaluation of machine translation", *Proceedings of ACL02*, pages 311–318.
11. Popović, M., Burchardt, A. (2011). "From Human to Automatic Error Classification for Machine Translation Output", *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 265–272, Leuven, Belgium.
12. Soricut, R., Echihab, A. (2010) "Trustrank: Inducing trust in automatic translations via ranking", *Proceedings of ACL11*, pages 612–621, Uppsala, Sweden.
13. Specia, L., Turchi, M., Cancedda, N., Dymetman, M., Cristianini, N. (2009) "Estimating the Sentence-Level Quality of Machine Translation Systems", *Proceedings of EAMT09*, pages 28–37, Barcelona, Spain.
14. Specia, L. (2011) "Exploiting objective annotations for measuring translation post-editing effort", *Proceedings of EAMT11*, pages 73–80, Leuven, Belgium.
15. Stymne, S., Ahrenberg, L. (2012). "On the practice of error analysis for machine translation evaluation", *LREC*, pages 1785-1790.
16. Vilar, D., Xu, J., D'Haro, LF., Ney, H. (2006) "Error Analysis of Machine Translation Output", *International Conference on Language Resources and Evaluation*, Genoa, Italy, pages 697–702.